The Elimination of Meaning in Computational Theories of Mind

[Published in *Reduction and Elimination in Philosophy and the Sciences*, Papers of the 31st International Wittgenstein Symposium, pp. 313-315, 2008. ISSN 1022-3398]

The Elimination of Meaning in Computational Theories of Mind

Paul Schweizer
School of Informatics
University of Edinburgh

Abstract: The traditional conception of the mind holds that semantical content is an essential feature distinguishing mental from non-mental systems. This traditional conception has been incorporated into the foundations of recent computational theories of mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. But a fundamental tension is then built into the picture - to the extent that symbolic 'representations' are formal elements of computation, their alleged content is completely gratuitous. Computation is a series of manipulations performed on *uninterpreted* syntax, and formal structure alone is sufficient for all effective procedures. I argue that the computational paradigm is thematically inconsistent with the search for content or its supposed vehicles. Instead, computational models of cognition should be concerned only with the *processing structures* that yield the right kinds of input/output profiles, and with how these structures can be implemented in the brain.

1. The Computational Paradigm

According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. For example, in the scholastic tradition revived by Brentano (1874), the *essential* feature of mental states is their 'aboutness' or intrinsic representational aspect. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. For example, in classical (e.g. Fodorian) cognitive science, Brentano's legacy is preserved in the view that the properly cognitive level is distinguished precisely by appeal to representational content. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars, and according to Fodor (1975), it is only when the states of a system are treated as representations that we are dealing with the genuinely cognitive level.

The classical paradigm in cognitive science derives from Turing's basic model of computation as rule governed transformations on a set of syntactical elements, and it has taken perhaps its most literal form of expression in terms of Fodor's Language of Thought hypothesis (LOT), wherein mental processes are explicitly viewed as formal operations on a linguistically structured system of internal symbols. But a fundamental tension is already built into the classical picture: a central purpose of the symbolic structures is to carry content, and yet, to the extent that they are formal elements of computation, their alleged content is completely gratuitous. Computation is essentially a series of manipulations performed on *uninterpreted* syntax, and formal structure alone is

sufficient for all effective procedures. The specification and operation of such procedures makes no reference whatever to the intended meaning of the symbols involved. Indeed, it is precisely this limitation to syntactic *form* that has enabled computation to emerge as a mathematically rigorous discipline. If syntax alone is not sufficient, and additional understanding or interpretation is required, then the procedure in question is, by definition, not an effective one. But then the purported content of mental 'representations' is rendered superfluous to the computations that comprise the 'cognitive' processes of cognitive science. The intended interpretation of internal syntax makes absolutely no difference to the formal mechanics of mind.

For a number of years now there has been a high profile struggle between opposing camps within the computational approach to the mind. In contrast to the classical paradigm derived from Turing, connectionist systems are based on networks of large numbers of simple but highly interconnected units that are brain-like in inspiration. But according to Fodor, the brain-like architecture of connectionist networks tells us nothing about their suitability as models of *cognitive* processing, since it still leaves open the question of whether the mind is such a network at the representational level. So a number of connectionists have taken up the challenge and seek out ways of projecting representational content onto artificial neural networks. One comparatively recent such attempt (Churchland, P.M.1998, Laakso, A. and G. Cottrell 2000, O'Brien, G. and J. Opie 2001) uses cluster analysis to locate 'vehicles' of representational content within artificial neural networks, where such clusters serve as surrogates for the classical notion of internal syntax.

However, I would contend that such attempts suffer from exactly the same built-in tension that afflicts the LOT model; namely, the purported content for which the clusters serve as vehicles does no work in the processing path leading from inputs to outputs. Just as in the classical case, the postulation of content within the connectionist framework is gratuitous, because it plays no role in the cognitive manipulation of inputs to yield the salient outputs. Indeed, if content weren't gratuitous, then computational versions of cognitive processing would be lamentably deficient in terms of their specification of the inputs. These are characterized solely in formal or syntactical terms, and content is entirely absent from the external stimuli recognized by the operations that can be defined within the model. If representational content were at all relevant, then cognitive systems would have to process content *itself*. But according to computational methods, content is not specified with the input, nor does it play any efficacious role in internal processing. So, from a perspective that takes computation as the theoretical foundation for cognition, it seems quite retrograde to posit content on top of the factors that do the actual work. Surely this is an exemplary occasion for invoking Ockham's razor.

2. Searle's Objection

Of course, John Searle's (1980) celebrated Chinese Room Argument (henceforward CRA) runs the dialectic in exactly the reverse direction: rather than taking the formal, syntactic nature of computation as a reason for eschewing content in a properly naturalistic approach to the mind, Searle instead takes it as a reason for rejecting computation as the appropriate theory of the mental. So, from the perspective of the present discussion, it is instructive to explicitly cast Searle's argument in terms of the

separability of syntactical structure from its intended meaning. In what follows I will abstract away from the somewhat picturesque details of Searle's original version and express the logical core of the CRA via two premises and a conclusion:

(1) semantical content is an essential feature of the mind,

(2) syntactical manipulations cannot capture this content, therefore

(3) the mind cannot be reduced to a system of syntactical manipulations.

Preimse (1) is an expression of the traditional conception of the mind, and is accepted by both Searle and by his opponents in orthodox cognitive science and AI. Classical cognitive science and AI view the mind according to the model of rule governed symbol manipulation, and premise (1) is embraced insofar as the manipulated symbols are supposed to possess representational content. Searle's dispute with cognitive science and AI centers on his rejection of the idea that internal computation can shed any real light on mental content, which leads to his conclusion (3), and a concomitant dismissal of the research paradigm central to cognitive science and AI.

In response, a standard line for defenders of this paradigm is to try and defuse the CRA by arguing against premise (2), and claiming that the manipulated symbols really do possess some canonical meaning or privileged interpretation. However, I would urge that this is a serious strategic error for those who wish to defend the computational approach. As stated above, a distinguishing mathematical virtue of computational systems is precisely the fact that the formal calculus can be executed without any appeal to meaning. Not only is an interpretation intrinsically unnecessary to the operation of computational procedures, but furthermore, there is no unique interpretation determined by the computational syntax, and in general there are arbitrarily many distinct models for any given formal system.

Computational formalisms are syntactically closed systems, and in this regard it is fitting to view them in narrow or solipsistic terms. They are, by their very nature, independent of the 'external world' of their intended meaning and, as mentioned above, they are incapable of capturing a unique interpretation, since they cannot distinguish between any number of alternative models. This can be encapsulated in the observation that the relation between syntax and semantics is fundamentally *one-to-many*; any given formal system will have arbitrarily many different interpretations. And this intrinsically one-to-many character obviates the possibility of deriving or even attributing a unique semantical content merely on the basis of computational structure.

The inherent limitations of syntactical methods would seem to cast a rather deflationary light on the project of explicating *mental content* within a computational framework. Indeed, they would seem to render hopeless such goals as providing a computational account of natural language semantics or propositional attitude states. Non-standard models exit even for such rigorously defined domains as first-order arithmetic and fully axiomatized geometry. And if the precise, artificial system of first-order arithmetic cannot even impose isomorphism on its various models, how then could a *program*, designed to process a specific natural language, say Chinese, supply a basis for the claim that the units of Chinese syntax posses a *unique* meaning?

So I think that the advocates of computation make the wrong move by accepting Searle's bait and taking on board the attendant 'symbol grounding problem' endemic to computational theories of mind. Instead I would accept Searle's negative premise (2) and

agree that computation is too weak to underwrite any interesting version of (1). Hence I would concur with Searle's reasoning to the extent of accepting the salient *conditional* claim that *if* (1) is true *then* (3) is true as well. So the real crux of the issue lies in the truth-value of (1), without which the consequent of the *if-then* statement cannot be detached as a free-standing conclusion. Only by accepting the traditional, *a priori* notion of mentality assumed in premise (1), does (3) follow from the truth of (2). And it's here that I diverge from the views of both Searle and orthodox cognitive science.

3. Representation as Heuristics

There have been a number of prominent positions advanced in negative reaction to 'classical' cognitive science that take anti-representationalism as one their hallmarks, including dynamical systems theory (e.g Van Gelder 1996), behaviour based robotics (e.g. Brooks 1991), approaches utilizing sensory-motor affordances (e.g. Noe 2004), who campaign on the platform of 'intelligence without representation'. In order to locate my position on the philosophical landscape, it is salient to note that it is *not* anti-representational in this sense. On my view, there could well be internal structures that play many of the roles that people would ordinarily expect of representations, and this is especially true at the level of perception, sensory-motor control and navigation – things like spatial encodings, somatic emulators, internal mirrorings of salient aspects of the external environment. So, unlike the anti-representationalists, I do not deny that there may be internal structures and stand-ins that various people would be tempted to *call* 'representations'.

But I would argue that this label should be construed in a weak, operational sense, and should not be conflated with the more robust traditional conception. To the extent that internal structures can encode, mirror or model external objects and states of affairs, they do so via their own causal and/or syntactic properties. And again, to the extent that they influence behaviour or the internal processing of inputs to yield outputs, they do this solely in virtue of their causal and/or syntactic properties. There is nothing about these internal structures that could support Searle's or Brentano's notion of original intentionality, and there is no independent or objective fact of the matter regarding their 'real' content or meaning.

The crucial point to notice is that these internal 'representations' do all their scientifically tangible *cognitive* work solely in virtue of their physical/formal/mathematical structure. There is nothing about them, qua efficacious elements of internal processing, that is 'about' anything else. Content is not an explicit component of the input, nor is it acted upon or transformed via cognitive computations. All that is explicitly present and causally relevant are computational structure plus supporting physical mechanisms, which is exactly what one would expect from a naturalistic account. In order for cognitive structures to do their job, there is no need to posit some additional 'content', 'semantical value', or 'external referent'. Such representation talk may serve a useful heuristic role, but it remains a conventional, observer-relative ascription, and accordingly there's no independent fact of the matter, and so there isn't a sense in which it's possible to go wrong or be mistaken about what an internal configuration is 'really' about. Instead, representational content is projected onto an internal structure when this plays an opportune role in characterizing the overall

processing activities which govern the system's interactions with its environment, and hence in predicting its salient input/output patterns. But it is simply a matter of convenience, convention and choice.

From the point of view of the system, these internal structures are manipulated directly, and the notion that they are 'directed towards' something else plays no role in the pathways leading from cognitive inputs to intelligent outputs. Hence the symbol grounding problem is a red herring – it isn't necessary to quest after some elusive and mysterious layer of content, for which these internal structures serve as the syntactic 'vehicle'. Syntactical and physical processes are all we have, and their efficacy is not affected by the presence or absence of meaning. I would argue that the computational paradigm is thematically inconsistent with the search for content or its supposed vehicles. Instead, computational models of cognition should be concerned only with the *processing structures* that yield the right kinds of input/output profiles, and with how such structures can be implemented in the brain. These are the factors that do the work and are sufficient to explain all of the empirical data, and they do this using the normal theoretical resources of natural science. Indeed, the postulation of content as the essential feature distinguishing mental from non-mental systems should be seen as the last remaining vestige of Cartesian dualism, and, contra Fodor, naturalized cognition has no place for a semantical 'ghost in the machine'. When it comes to computation and content, only the vehicle is required, not the excess baggage.

Literature:

Brentano, F. 1874 *Psychology from an Empirical Standpoint*.

Fodor, J. 1975 *The Language of Thought.* Harvester Press.

Churchland, P.M. 1998 "Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered", *Journal of Philosophy*, 96(1): 5-32.

Laakso, A. and G. Cottrell 2000 "Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems", *Philosophical Psychology*, 13(1): 47-76.

O'Brien, G. and J. Opie 2001 "Connectionist Vehicles, Structural Resemblance, and the Phenomenal Mind", *Communication and Cognition*, 34: 13-38.

Searle, J. 1980 "Minds, Brains and Programs", *Behavioral and Brain Sciences*, 3: 417-424.

Van Gelder, T. 1996 "Dynamics and Cognition" in *Mind Design II*, J. Haugeland (ed.), MIT Press.

Brooks, R. 1996 "Intelligence without Representation" in *Mind Design II*, J. Haugeland (ed.), MIT Press.

Noë, A. 2004 *Action in Perception*, MIT Press.